Lending Club Case Study

Understanding and Predicting Loan Defaults

Abinash Rout & Abhishek Tomar

26th June 2024

Introduction

- Importance of predicting loan defaults for a consumer finance company
- Reduce financial losses and improve loan approval processes

Business Problem

- Risk of financial loss due to loan defaults
- Need to identify high-risk loan applicants
- Dataset containing information on past loan applicants

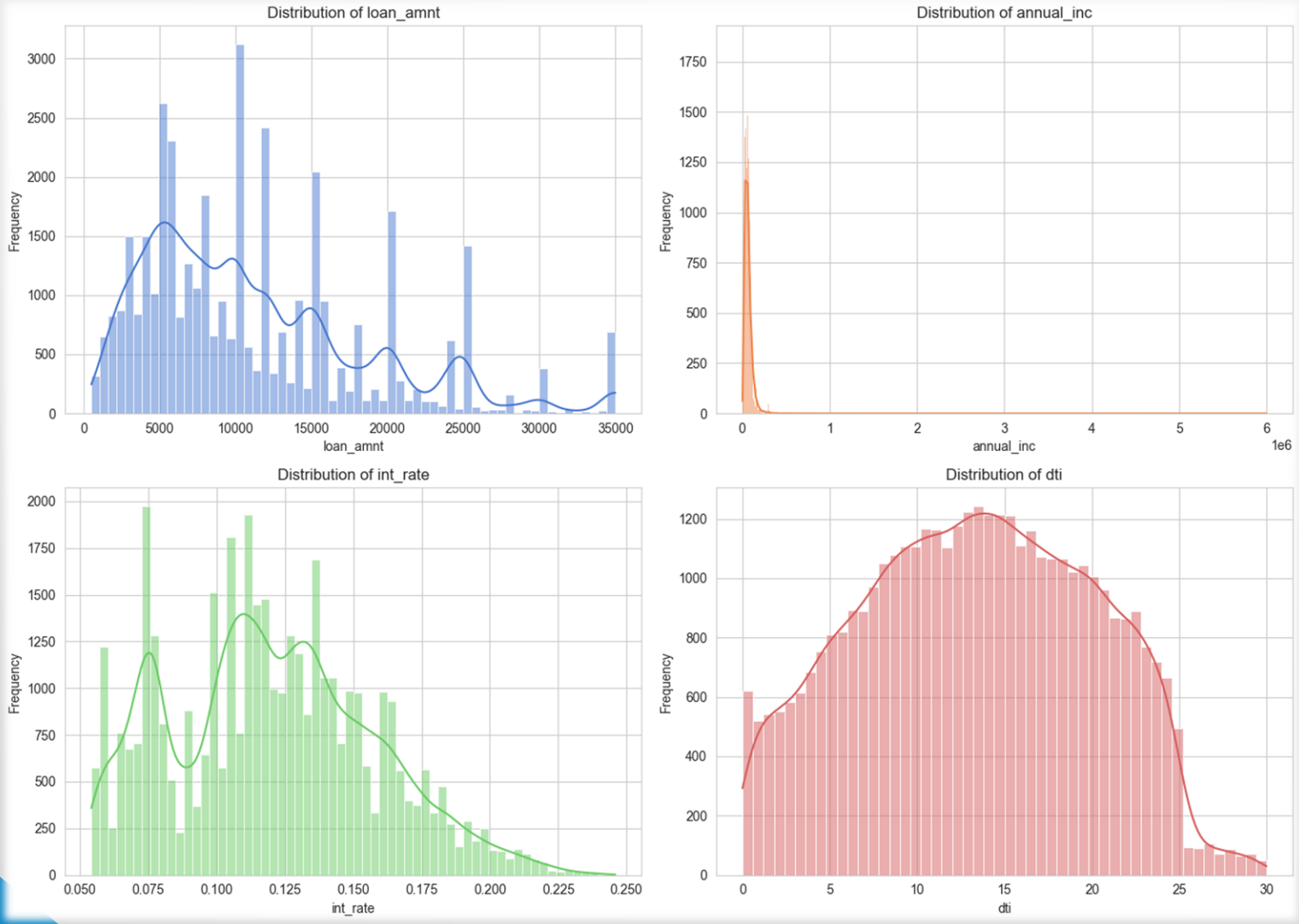
Data Overview

- Description of the dataset (2007-2011 loans)
- Key features: loan amount, interest rate, term, grade, purpose, etc.
- Initial data shape: 39,717 entries and 111 features

Data Cleaning and Preprocessing

- Handling missing values by dropping columns with more than 50% missing values
- Imputing remaining missing values with appropriate methods
- Correcting data types (percentage, date)
- Dropping unnecessary columns (e.g., columns with only one value, unique identifiers)
- Dropping unnecessary columns
- Final dataset shape after cleaning

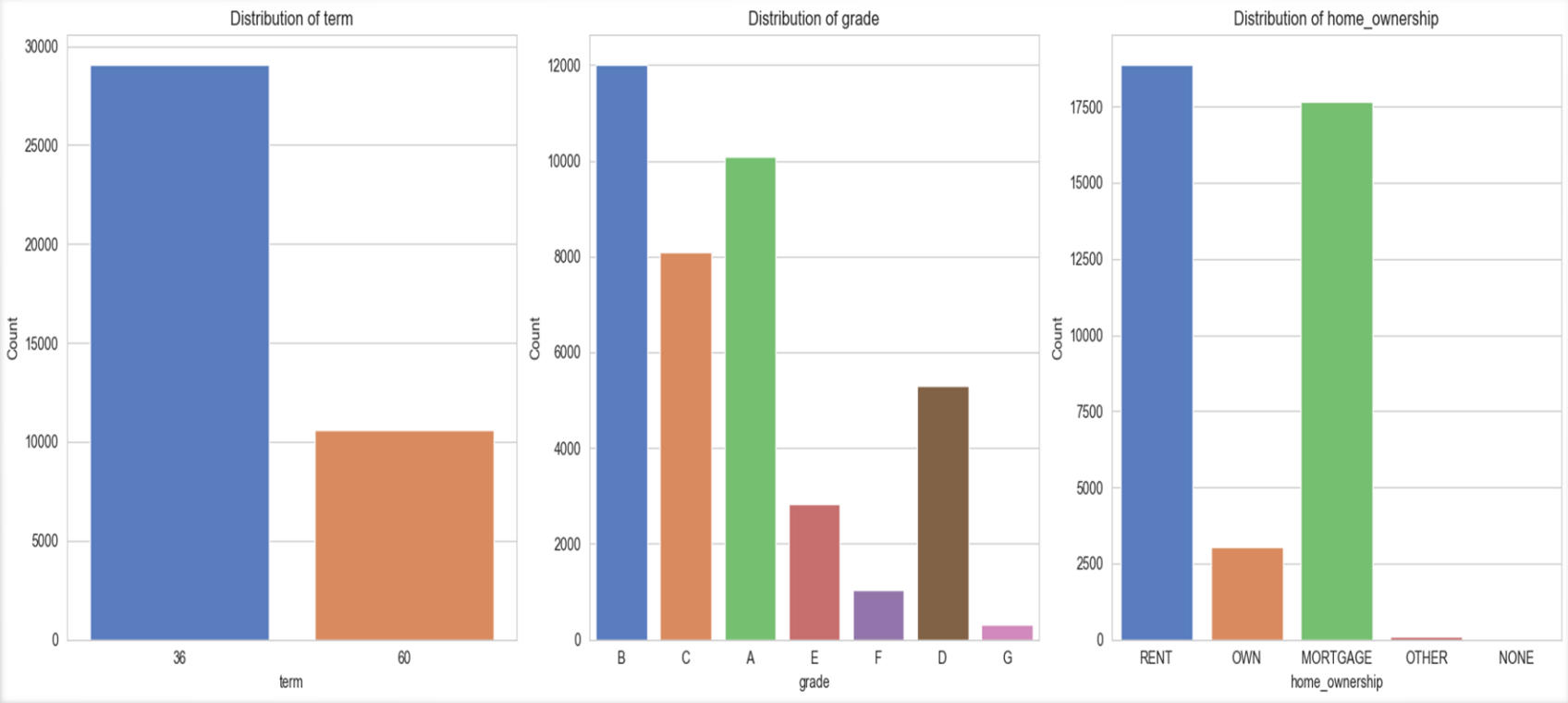
Univariate Analysis - Numerical Features



Univariate Analysis - Numerical Features Insights

- Distribution of loan amount: Most loans are in the lower range.
- Distribution of annual income: Right-skewed, with most incomes in the lower range.
- Distribution of interest rate: Most loans have interest rates around 10-15%.
- Distribution of debt-to-income ratio (DTI): Shows the distribution of borrowers' DTI.

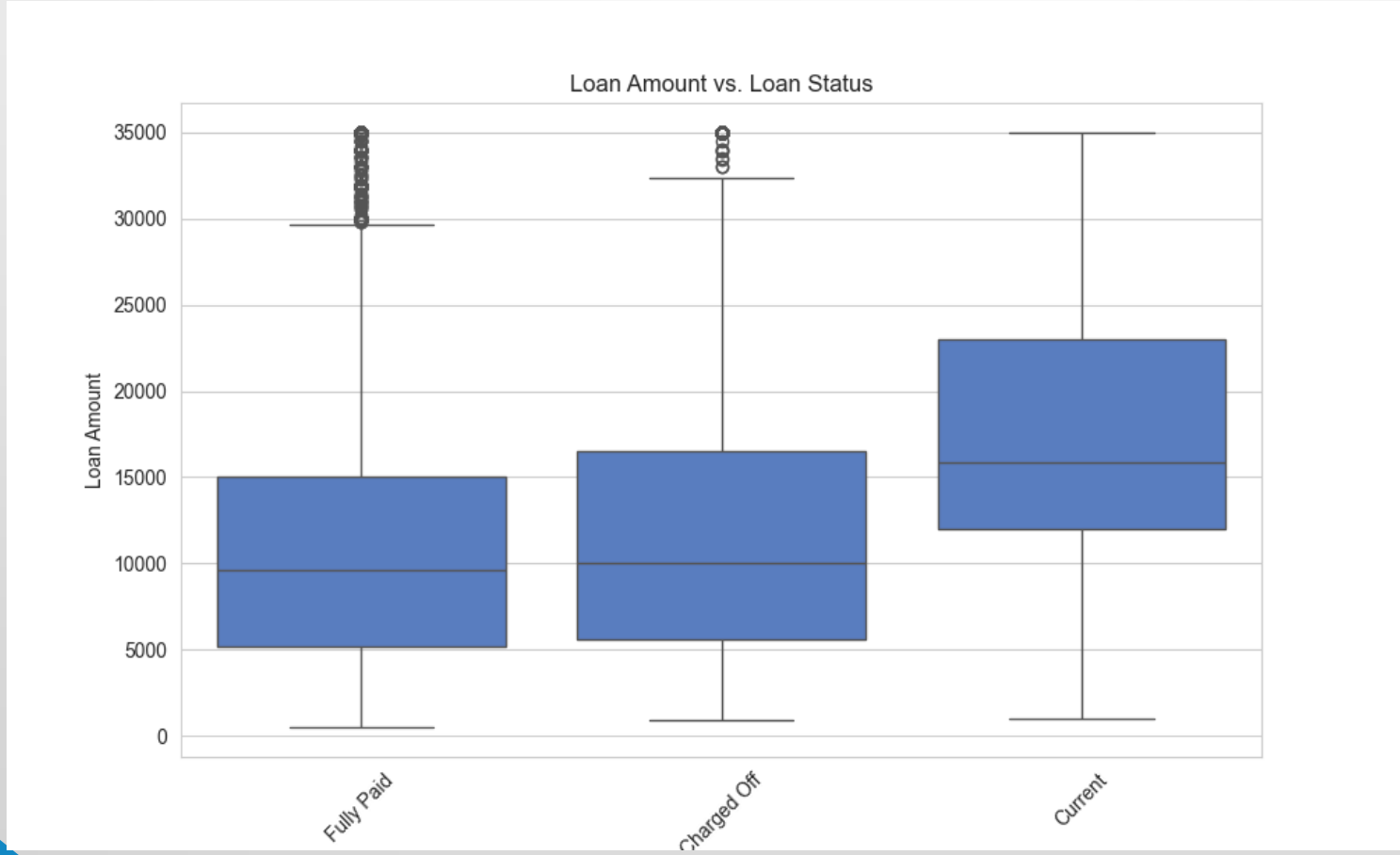
Univariate Analysis - Categorical Features



Univariate Analysis - Categorical Features Insights

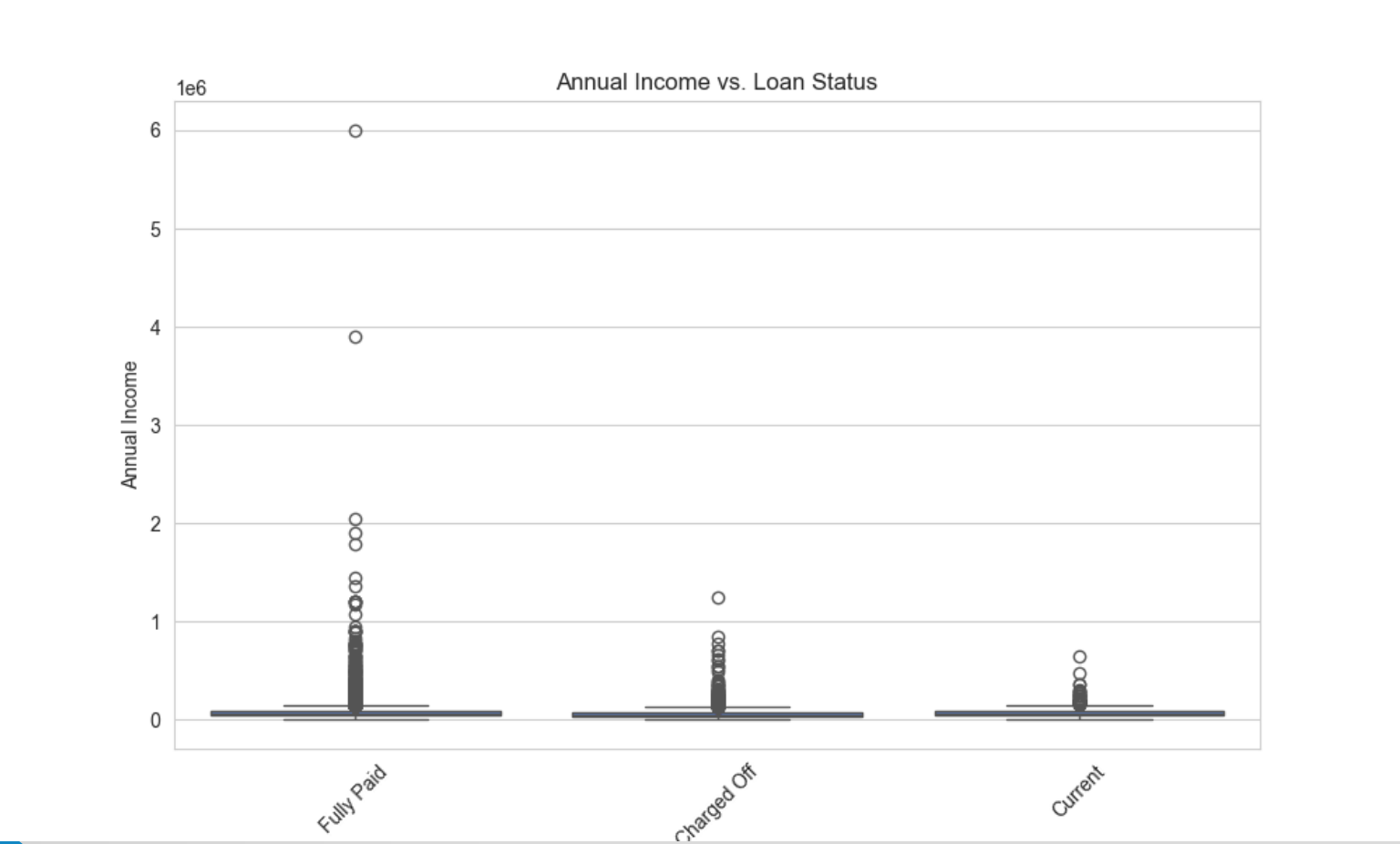
- Distribution of loan term: Majority of loans have a term of 36 months.
- Distribution of loan grade: Indicates varying levels of creditworthiness.
- Distribution of home ownership: Status of borrowers' home ownership.

Bivariate Analysis - Loan Amount vs. Loan Status



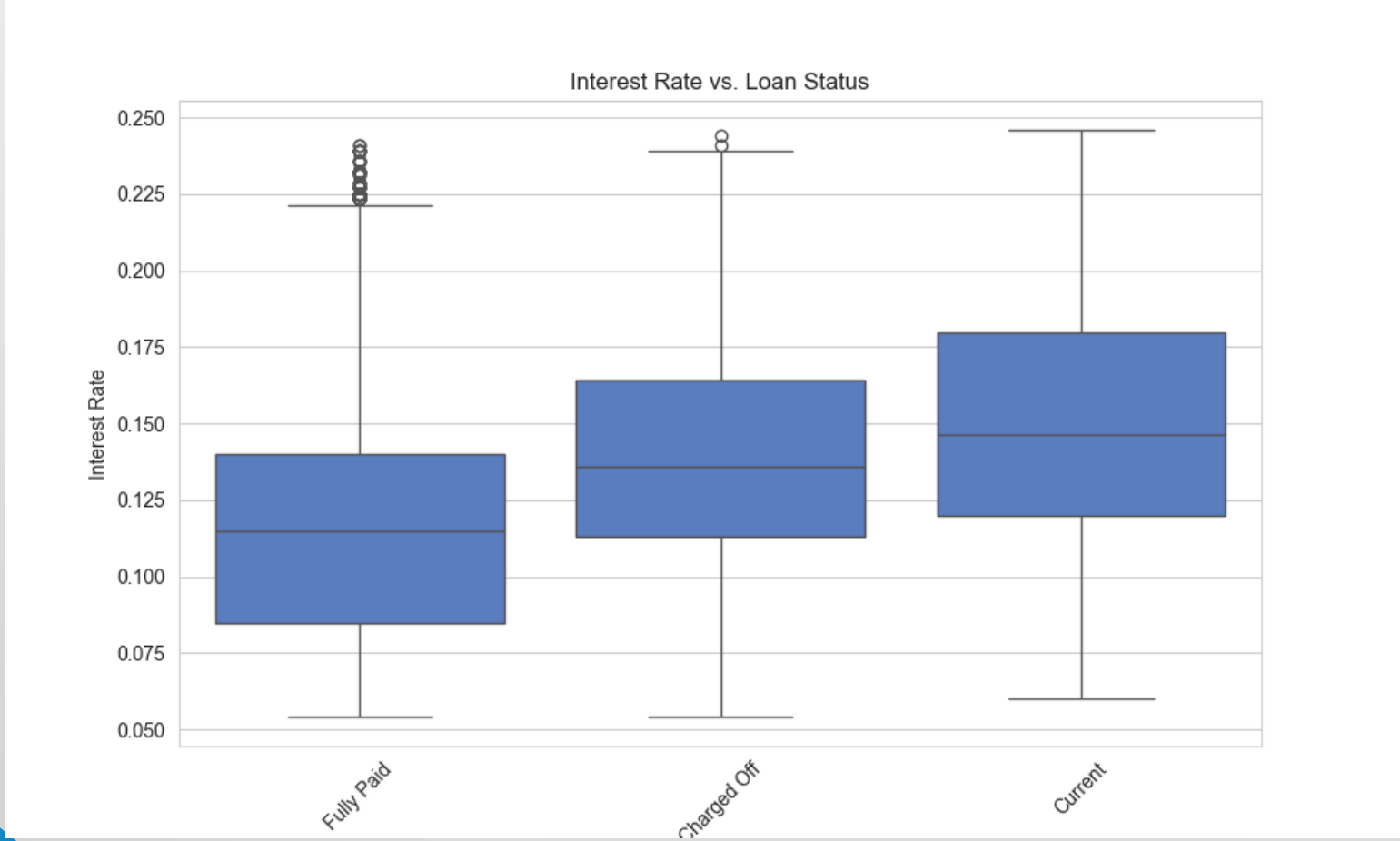
Higher loan amounts are associated with a higher risk of default.

Bivariate Analysis - Annual Income vs. Loan Status



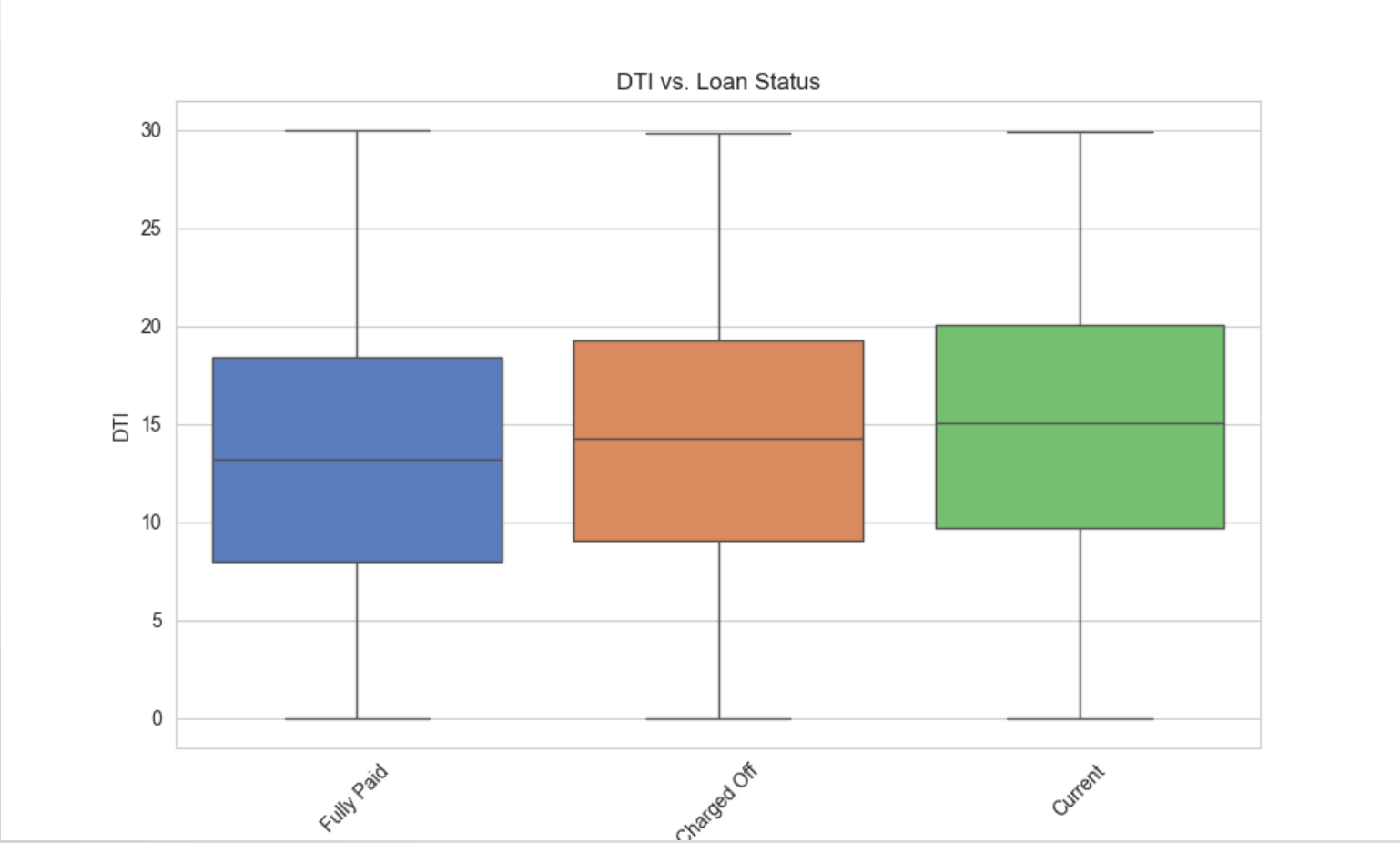
Annual income does not show a strong relationship with loan default status

Bivariate Analysis - Interest Rate vs. Loan Status



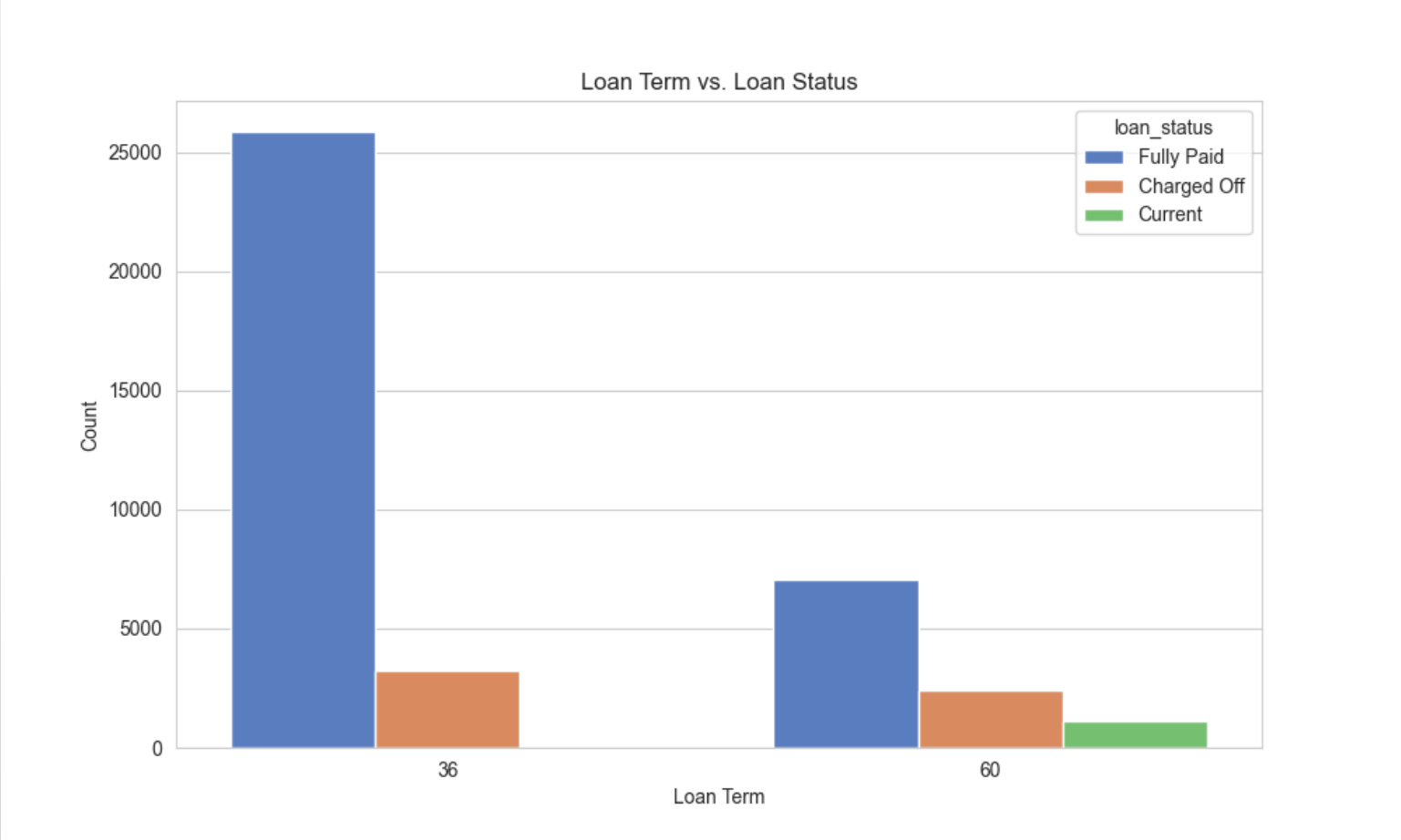
Higher interest rates are associated with a higher risk of default.

Bivariate Analysis - DTI vs. Loan Status



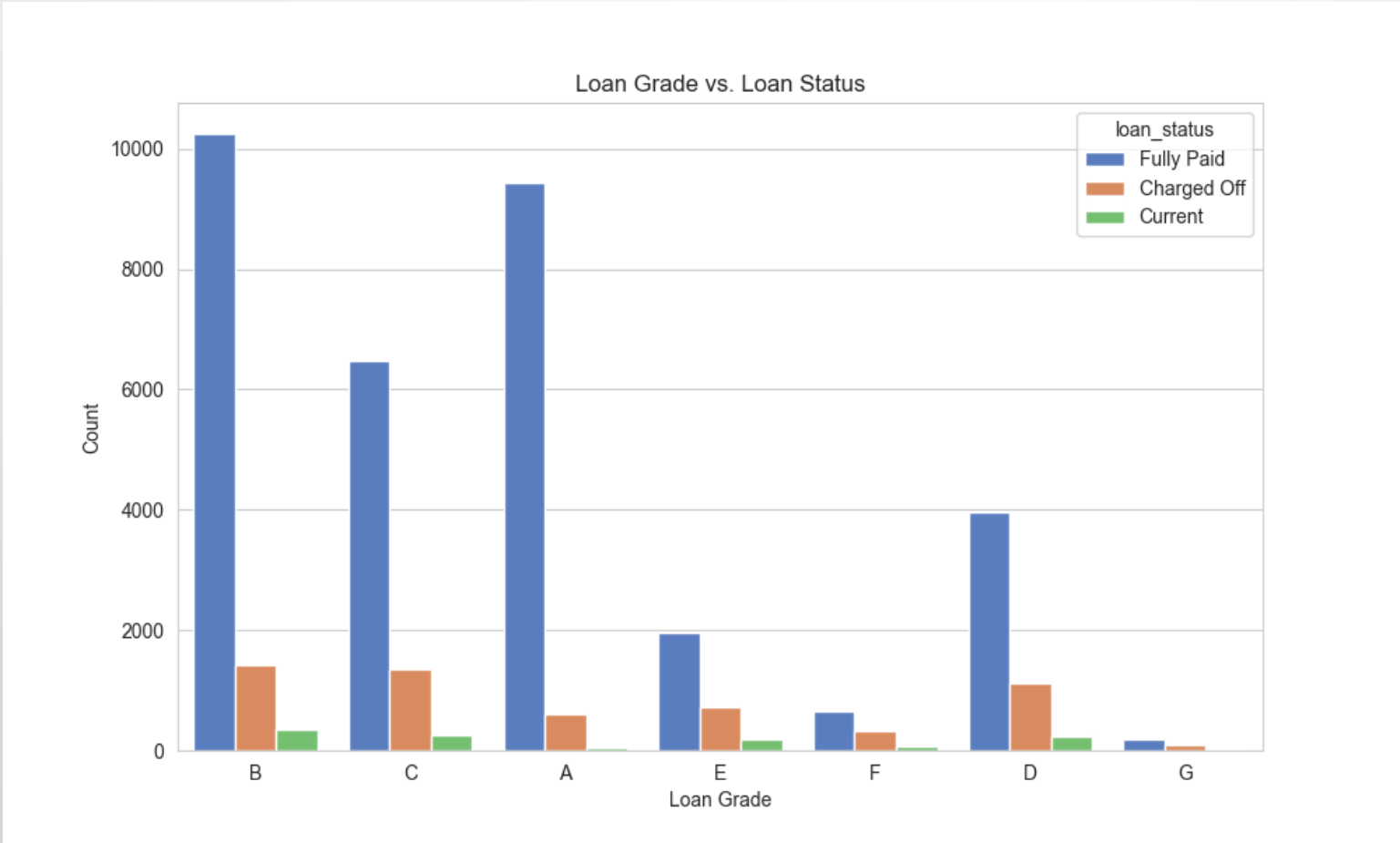
Higher debt-to-income ratios are associated with a higher risk of default.

Bivariate Analysis - Term vs. Loan Status



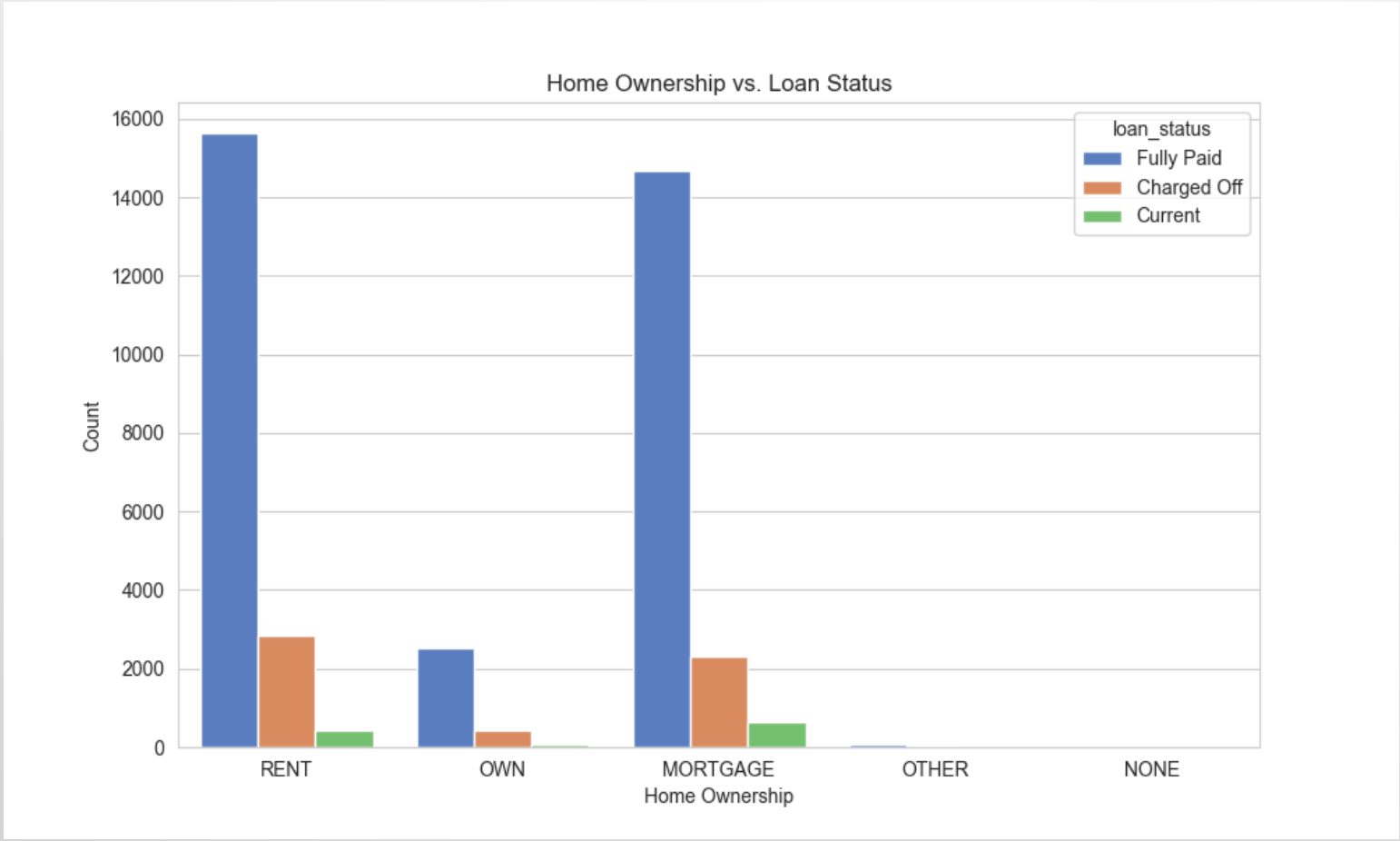
Loans with longer terms (60 months) are associated with a higher risk of default.

Bivariate Analysis - Grade vs. Loan Status



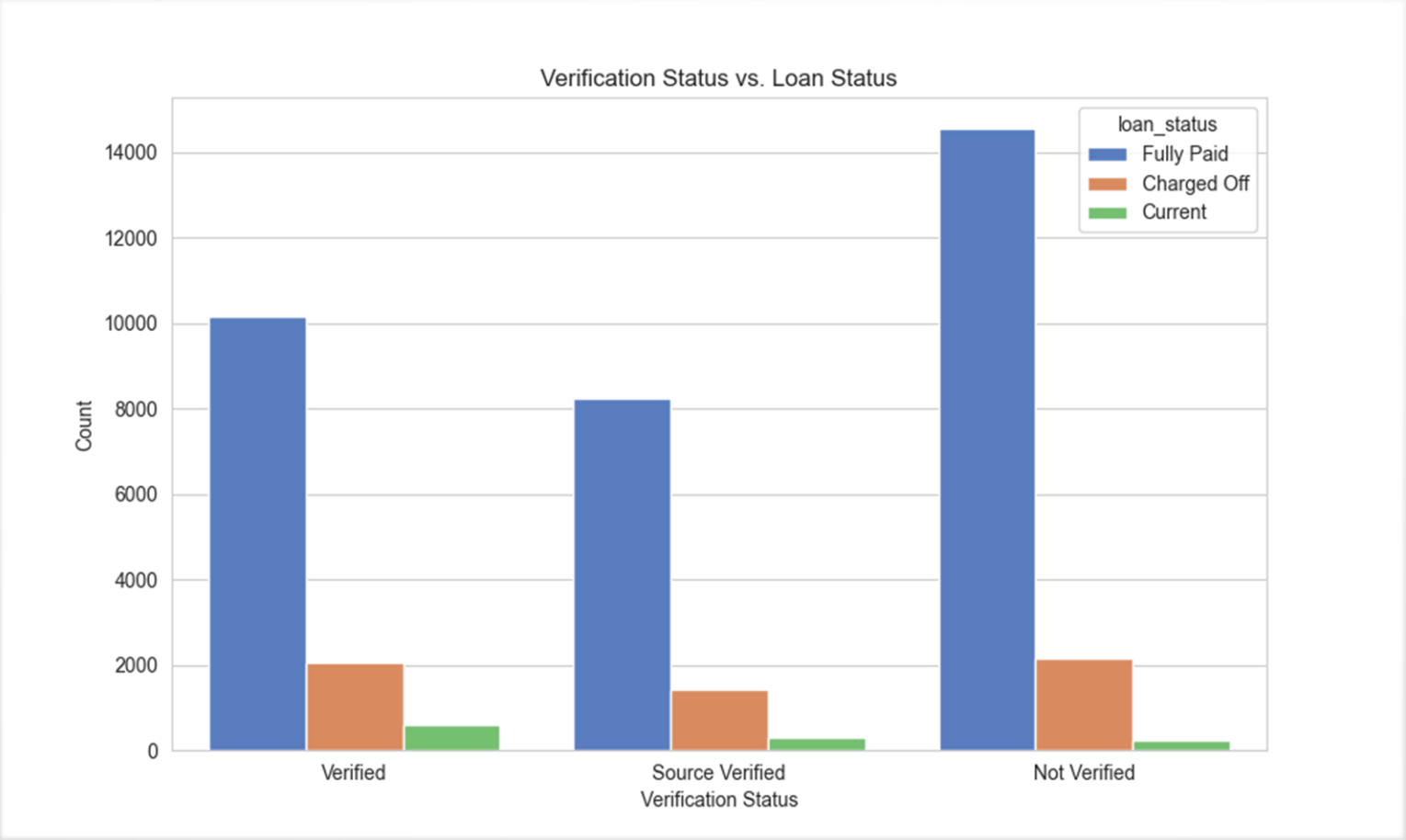
Lower loan grades (D, E, F, G) have a higher proportion of defaults compared to higher grades (A, B).

Bivariate Analysis - Home Ownership vs. Loan Status



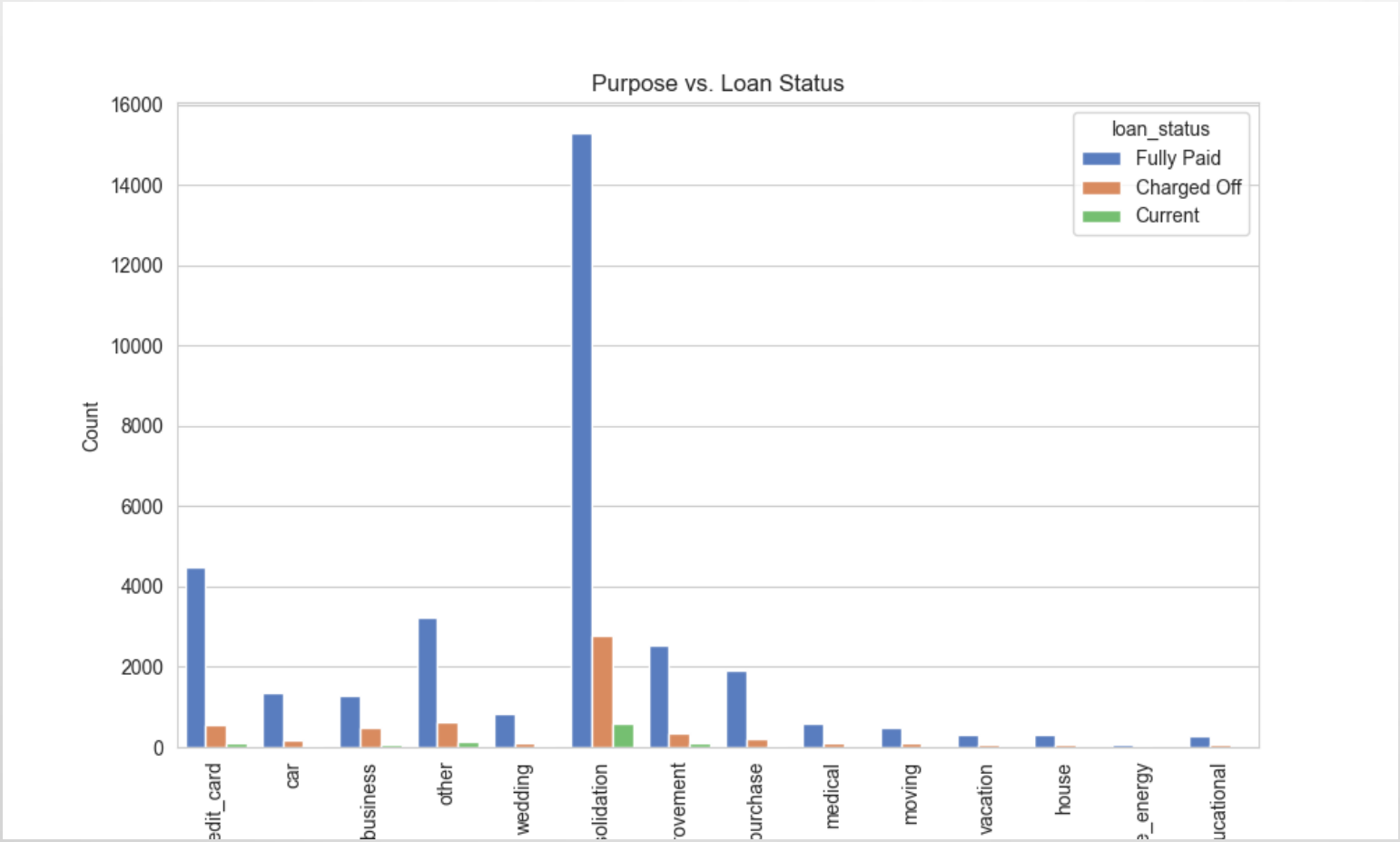
Home ownership status does not significantly impact the likelihood of default.

Bivariate Analysis - Verification Status vs. Loan Status



Verification status has some impact on loan default risk, with verified loans showing a slightly lower default rate.

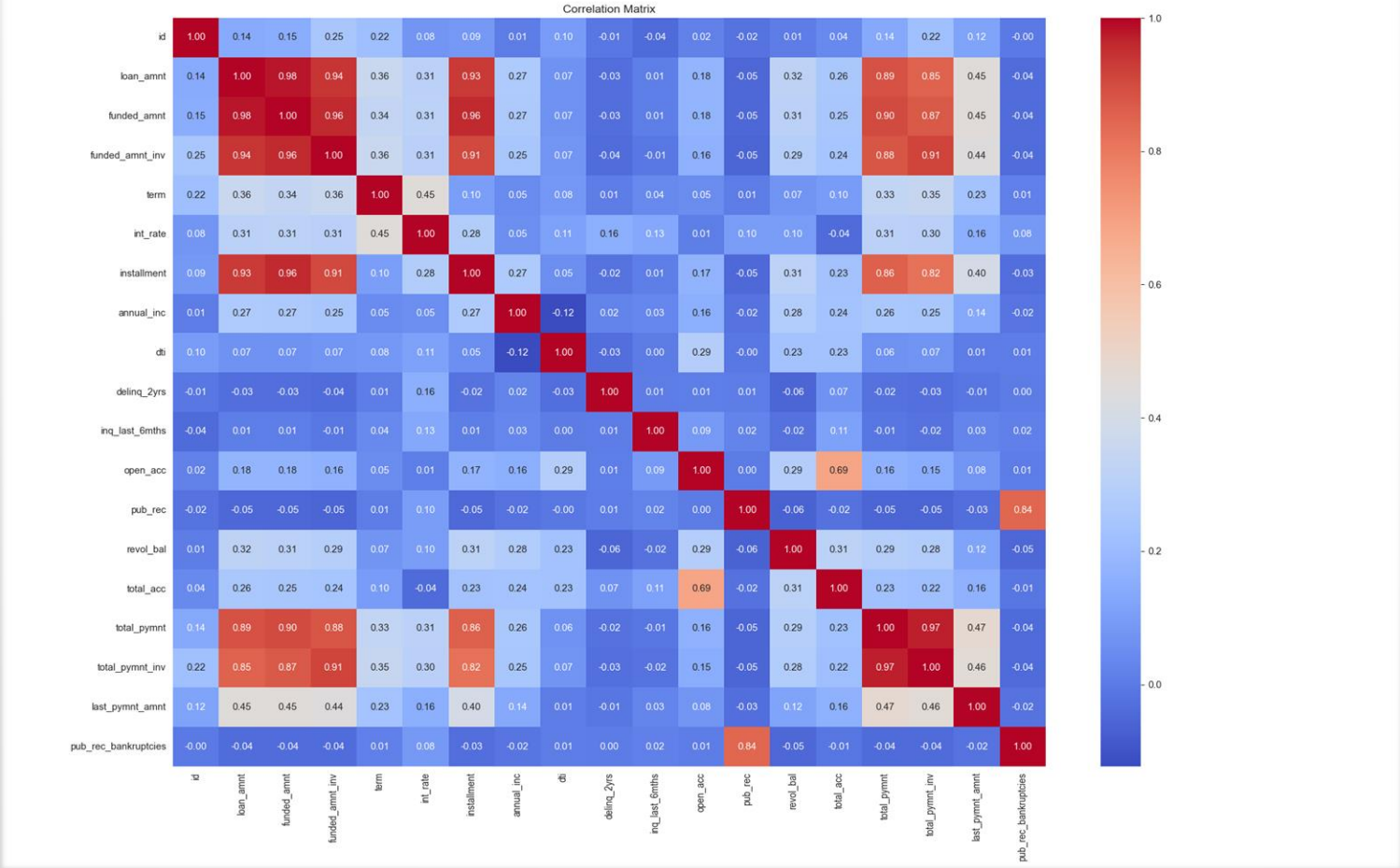
Bivariate Analysis - Purpose vs. Loan Status



Bivariate Analysis - Purpose vs. Loan Status Insights

- Loans for credit card refinancing and debt consolidation have a higher proportion of fully paid loans.
- Loans for small business, renewable energy, and educational expenses show a higher proportion of defaults.

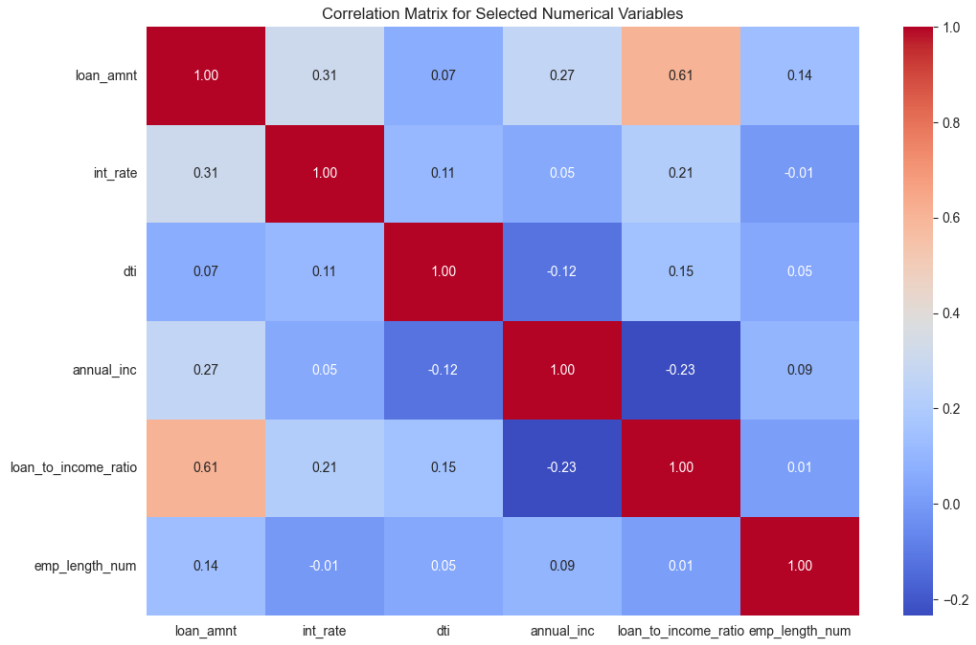
Correlation Analysis



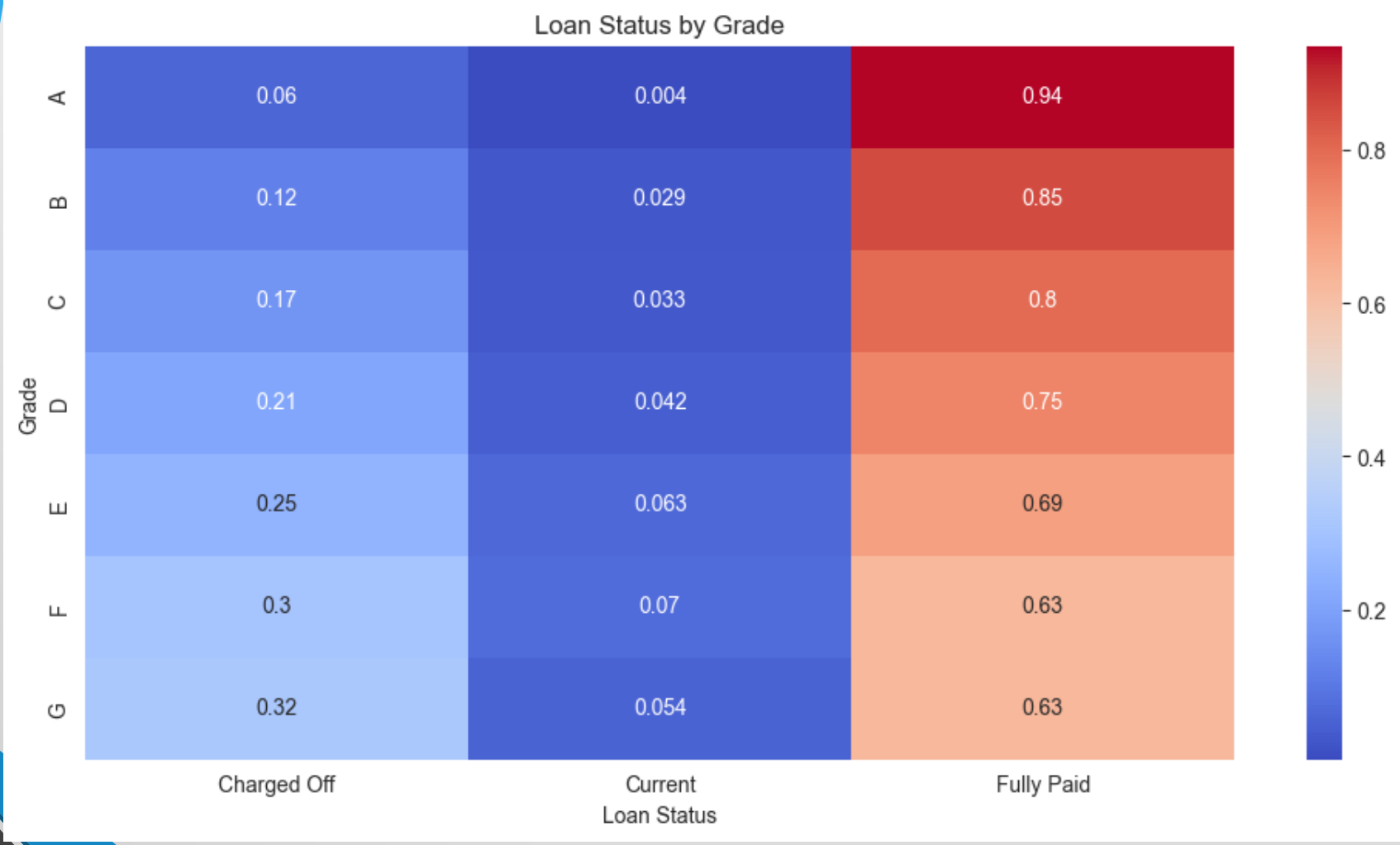
Correlation Analysis Insights

- High correlation between loan amount and installment.
- High correlation among total payment metrics (total_pymnt, total_rec_prncp, total_pymnt_inv).
- Moderate correlation between interest rate and loan amount/installment.
- Low correlation between annual income and other numerical features.

Multivariate Analysis - Correlation Matrix for Selected Numerical Variable



Multivariate Analysis – Loan Status by Grade

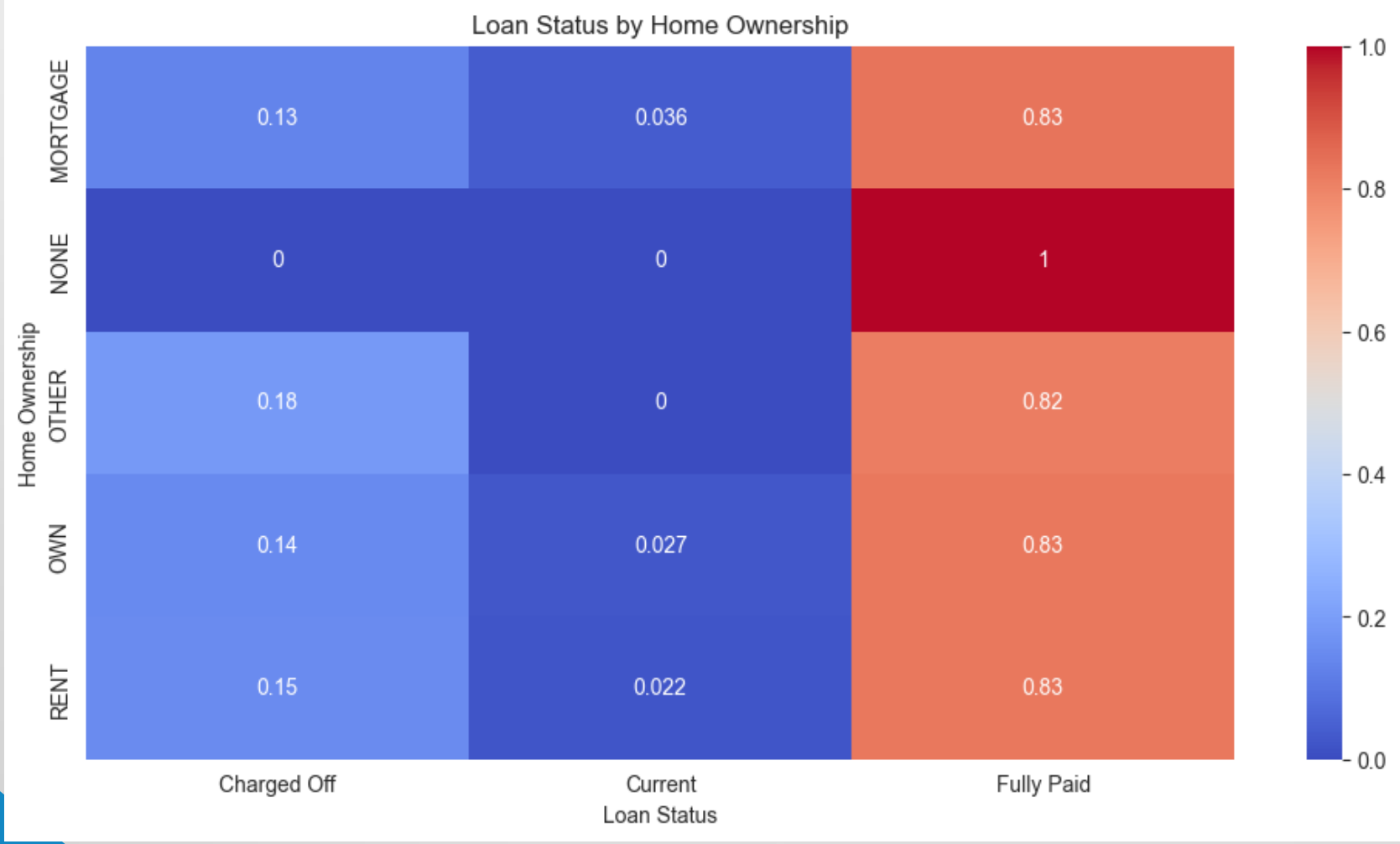


Multivariate Analysis – Grade and Loan Status Heatmap Insights

The heatmap for loan status by grade shows the distribution of loan statuses across different loan grades:

- Higher grades (A, B) have a larger proportion of fully paid loans.
- Lower grades (D, E, F, G) have a higher proportion of charged-off (defaulted) loans.
- This indicates that loan grade is a strong indicator of loan default risk.

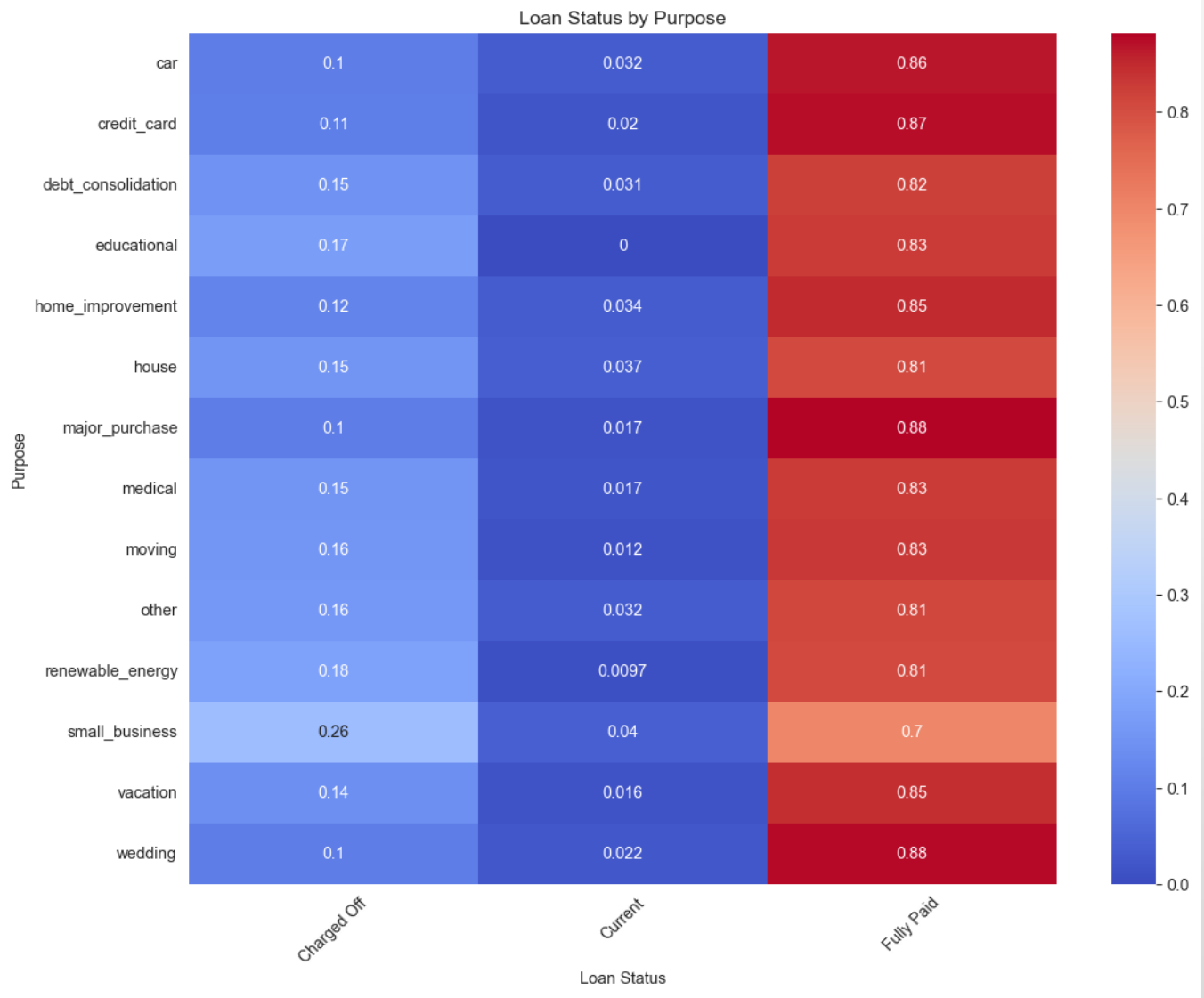
Multivariate Analysis – Loan Status by Home Ownership



Multivariate Analysis – Loan Status and Home Ownership Heatmap Insights

- The home ownership heatmap reveals that applicants who own their homes tend to have a higher proportion of fully paid loans.
- Applicants who rent have a slightly higher proportion of charged-off loans compared to those who own or have a mortgage.

Multivariate Analysis – Loan Status by Purpose



Multivariate Analysis – Loan Status and Purpose Heatmap Insights

- The purpose heatmap shows that loans for purposes like credit card refinancing and debt consolidation have a higher proportion of fully paid loans.
- Loans for purposes such as small business, renewable energy, and educational expenses show a higher proportion of charged-off loans.

Final Conclusions

- High-risk indicators: Loan amount, interest rate, DTI(Debt-to-Income Ratio), loan grade
- Moderate impact indicators: Loan term, purpose
- Low impact indicators: Annual income, home ownership
- Business implications: Risk management strategies, refining loan approval criteria, targeted interventions for high-risk applicants